Using Bidirectional Attention Flow & Option Pointer for Question Answering

Mei Fu, Yuqing Xia

Introduction

- Task

Use deep learning models to help machine better understanding relations between context and question, including vector representation and similarity matrix representation.

- Dataset

CNN/DailyMail dataset released by DeepMind

1) anonymized the noun entities (e.g. @entity01) to force the model to learn from the context instead of the entity itself

2) the news stories provide more sufficient background information compared with other dataset such as SQuAD, which normally contains only several sentences

Story	@entity0 , @entity1 (@entity2) @entity3 lure @entity5 and @entity6 migrants by offering discounts to get onto overcrowded ships if people bring more potential passengers , a @entity2 investigation has revealed . a smuggler in the @entity1 capital of @entity0 laid bare the system for loading boats with poor and desperate refugees , during a conversation that a @entity2 producer secretly filmed ……
Query	@placeholder investigation uncovers the business inside a human smuggling ring
Answer	@entity2
Words	@entity3:Smugglers @entity2:CNN @entity1:Libyan @entity0:Tripoli ……

Baseline - Logistic Regression

1. Entity Frequency

- words that be chosen as target entity has average of 8 times of appearance in the context

2. First Index Location

- content that appear before are more important as word that appear later
- 80% of entities show up in the 1/3 of the whole contexts

3. Bi-gram Exact Match

- tokens at the end or around (2 indexes around) the matched results have high probabilities as correct answers

4. Distance

- minimum distance between each entities in current news set each entities or clues in the question data
- entities that have distance between 1 - 10 tend be a correct answer

5. Embedding Similarity

- within the top 5 similar words of @placeholder

Stronger Models!

VECTOR REPRESENTATION

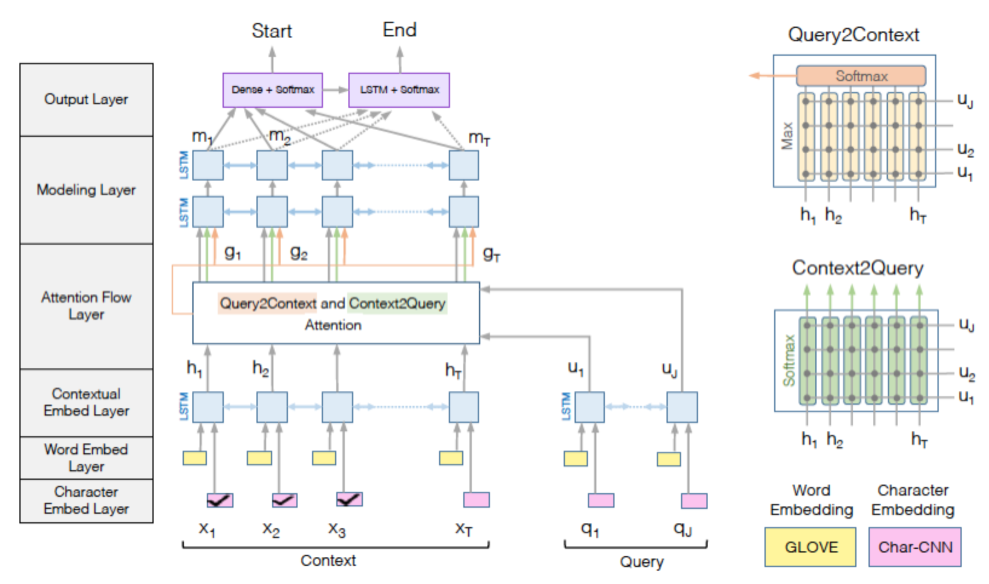
- Word Embedding
- Char Embedding (average)

INPUT

- Passage Input (n,300,)
- Question Input (n,46,)
- Option Input (n,102)

OUTPUT

- Option Input (n,102)



Minjoon Seo, Aniruddha Kembhavi,Ali Farhadi, Hananneh Hajishirzi, 2017. Bidirectional attention flow for machine comprehension. ICLR.

Challenges & Solutions

1. **Dynamic entity representation**

    LSTM Encoding + CNN char embedding with max-pooling

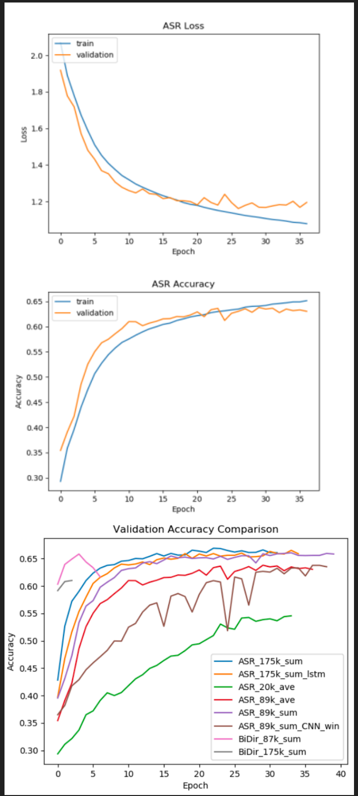
2. **Context Window**

    19 word window of that is centered on every entity

3. **Sum entity probabilities in multiple locations**

    Adding a masking layer which takes all word codes of entity as input

Results



Model (Train Size) \ Accuracy	**Train**	**Validation**	**Test**
Logistic Regression (87k)	-	38.0	41.4
BiGRU + Attention x 2 (20k) (simplified R-Net)	46.0	-	-
BiGRU + Attention + Masking (ave) (20k)	60.5	54.6	54.9
BiGRU + Attention + Masking (ave) (87k)	64.1	63.8	63.6
BiGRU + Attention + Masking (sum) (87k)	67.3	66.1	66.9
CNN +(win)+ BiGRU + Att + Masking(sum) (87k)	62.9	63.8	64.7
BiGRU + Attention + Masking (sum) (175k)	65.2	66.9	67.2
BiLSTM + Attention + Masking (sum) (175k)	66.3	66.5	67.1
Kadlec et al. (2016)[1]	-	68.6	69.5
Chen et al. (2016)[2]	-	73.8	73.6
Minjoon et al. (2017)[3]	-	76.3	76.9

1. Rudolf Kadlec, Martin Schmid, Ondrej Bajgar & Jan Kleindienst. IBM Watson. 2016. Text Understanding with the Attention Sum Reader Network. arXiv preprint arXiv:1603.01547v2.

Thanks!